图书情报工作
第65卷 第19期 2021年10月
ChinaXiv合作期刊

多节点多关系的混合网络社团划分研究综述*

■ 蒋璐[1,2] 陈云伟[1,2]
[1] 中国科学院成都文献情报中心科学计量与科技评价研究中心(SERC) 成都 610041
[2] 中国科学院大学经济与管理学院图书情报与档案管理系 北京 100190

摘 要:[目的/意义]旨在对多节点多关系混合网络社团划分方法进行梳理,探析现有社团划分方法存在的问题及面临的困难,预见未来的发展趋势。[方法/过程]对近年来有关多节点类型、多关系类型混合网络的社团划分方法研究工作进行系统梳理,从基于概率生成模型、元路径、种子节点、扩展模块度以及混合网络同构5个方面对划分方法进行阐述,归纳混合网络社团划分常用的评估指标:标准化互信息NMI、调整兰德指数ARI和模块度Q,指出社交媒体、学术网络、欺诈检测3个应用场景。[结果/结论]归纳出多节点多关系混合网络社团划分方法的适用性及优缺点,揭示了当前发展面临的挑战,为后续的混合网络分析研究提供新的视角,并展望今后可能进一步拓展的相关研究方向。

关键词:多节点 多关系 混合网络 社团 评估
分类号:G250
DOI:10.13266/j.issn.0252-3116.2021.19.014

chinaXiv:202304.00470v1

网络科学是近年来最活跃的热点研究领域之一[1],已被成功应用于诸多领域,如研究个人之间社会关系的社会科学[2]、研究基因和蛋白质之间相互作用的生物学[3]、研究大脑结构和功能的神经科学[4]等。已有研究发现,整个网络是由若干个"社团"组成的,社团(community)也称为集群或模块,是具有公共属性或在网络中具有相同角色的顶点集[5]。每个社团内部节点间的连接相对非常紧密,但是各个社团之间的连接相对稀疏[6-7]。识别网络的社团结构不仅可以揭示节点间的相似性,还可以揭示社团内部的工作原理,有助于理解网络结构特征和潜在语义信息。因此,社团划分被认为是理解和分析网络的一个基本手段[8],可以根据已观察到的网络结构信息及节点的属性信息来预测网络的社团结构[9],并将密集连接的节点聚集到社区中[5,10]。

目前,鲜有学者对多节点多关系混合网络的社团划分方法进行系统的梳理归纳,大多是对同构网络的社团划分方法进行综述[11-12]。李辉等[13]从模块度优化、标签传播、局部扩展、流式分析、深度学习方法5个方法角度对复杂网络中的社团划分方法进行综述,但其讨论的方法大多只适用于单节点单关系的网络;张瑞红等[14]针对单节点单关系、单节点多关系、多节点多关系混合网络的社团划分方法分别进行了梳理,但其仅对多节点多关系混合网络的部分社团划分方法进行梳理,归纳不够全面、不成体系,未明确指出其发展历程及优缺点。因此,本文聚焦多节点多关系混合网络,检索国内外相关文献,首先明确多节点类型、多关系类型混合网络的定义,并与其他网络进行概念辨析;其次介绍了多节点多关系混合网络的社团划分方法与主要评价指标,揭示每一类算法的发展过程、优缺点及适用性,以期让相关研究人员对该领域有更清晰和全面的认识,同时为社团划分算法的深入研究提供理论依据;最后阐述了其应用场景及发展面临的困难,为构建一个网络适用性强、复杂度低、结合网络拓扑结构与文本信息的社团划分算法提供思路,对今后的动态网络社团划分方法、混合网络标准评估指标的提出作出了展望。

1 多节点类型、多关系类型混合网络的定义

一直以来有关网络的定义层出不穷,现将各网络

* 本文系国家社会科学基金项目"用于科学结构分析的混合网络的社团研究"(项目编号:19XTQ012)研究成果之一。
作者简介:蒋璐(ORCID:0000-0002-5514-9860),硕士研究生;陈云伟(ORCID:0000-0002-6597-7416)),研究员,硕士生导师,通讯作者,E-mail:chenyw@clas.ac.cn。
收稿日期:2021-04-06 修回日期:2021-08-02 本文起止页码:142-150 本文责任编辑:易飞

定义表述如表 1 所示:

表 1　各网络的定义

网络	定义
复杂网络	具有自组织、自相似、吸引子、小世界、无标度中部分或全部性质的网络[15]
超网络	规模巨大、连接复杂、节点具有异构性的网络[16]
异构网络	对象类型满足|A| > 1、关系类型|R| > 1 的网络[17]
同构网络	与异构网络对应,仅包含一种节点类型和关系类型的网络[18]
混合网络	含有多种节点类型或多种关系类型的网络[14]

复杂网络和超网络两者所指向的网络较为宽泛,强调的是一种呈现高度复杂性的网络,并未对网络的节点和关系作具体说明。异构网络强调的是网络拓扑结构的复杂性,其更注重异构关系,对同类型对象之间的同构关系关注较少[19]。同构网络仅从一个视角反映某一方面的联系,在研究科学结构、识别研究前沿和技术机会上存在局限[20]。混合网络本质上属于异构网络的范畴,强调的是多种节点与多种关系的混合,充分考虑了异构关系和同构关系,体现出功能的丰富性。故本研究采用"混合网络"定义,以便在符合真实网络的情况下,将研究人员的焦点从网络拓扑结构的构建上聚焦到功能的提升上。

多节点类型、多关系类型的混合网络即是指包含多种节点类型和多种关系类型的网络,其特性主要表现在以下两个方面:①节点的多样性,即节点类型多,例如在学术网络中,节点可以为作者、文献、关键词、期刊等,在医疗网络中,节点可以为医生、药品、病人等;②关系的丰富性,在学术网络中,关系可以包含作者合作关系、作者引用关系、文献引用关系、作者与文献的隶属关系等,在医疗网络中,关系可以包含医生开药关系、病人服药关系等。多节点多关系混合网络能从多个视角整合多方面的关系,利用网络中多类型节点和链接的丰富语义,从相互关联的数据中发现丰富的知识,捕获真实世界中最根本的语义信息[21]。因此为全面地了解某个领域内的科学结构信息,深入研究多节点多关系混合网络是十分有必要的。

2　多节点多关系混合网络的社团划分方法

由于学者们对多节点多关系混合网络中的多类型节点在同一网络中的配置原则、多类型关系转换成数据关联规则的方法[22]、不同关系边的权重方案、分层网络不同层级间的社团划分方法均尚未达成共识,直接将传统的同构网络社团划分方法应用在混合网络中尚存在不足。因此目前对多节点多关系混合网络社团划分方法的研究多集中于以下两种:一种是扩展现有的算法来直接处理混合网络,另一种是将混合网络降维为同构网络再进行社团划分[23－25]。基于以上两种划分思路,多节点多关系混合网络的社团划分方法主要有以下 5 种类型:

2.1　基于概率生成模型的方法

基于概率生成模型的方法包括基于排序的方法和概率统计模型方法。

2.1.1　基于排序的方法

在基于概率生成模型的方法中,部分算法将排名问题与社团划分问题相结合,排名与社团划分是相辅相成的:好的排名增强社团划分结果,好的社区亦能改进排名[26]。RankClus[27]是最早提出的基于混合网络的排序聚类算法,但其只适用于两种类型的节点;Y. Z. Sun 等基于 RankClus 提出了一种新的算法 NetClus[28],利用多类型节点之间的链接来生成高质量的网络集群,有更好的聚类效果,但其仅适用于星型网络结构,且需要提前知晓在数据集中具有代表性的对象。由于以上算法均没有普适性,M. Ji 等提出 RankClass[29]算法,使其适用于任意网络模式的混合网络,且可充分利用任何数据对象的标签信息;赵焕对经典的 NetClus 算法进行改进,提出 MAO-NetClus 算法,针对 Web 服务、提供商、用户 3 个类型的节点及其之间的关系,实现了基于多节点多关系混合网络的 Web 服务聚类,设计了 Web 服务推荐系统原型[30]。此外,为实现动态混合网络的社团划分,揭示每种类型节点的演化过程,M. Gupta 等[31]提出 EnetClus 算法,该算法执行一种演化聚类,使用时间平滑方法显示随时间变化的聚类;C. H. Qiu 等[32]提出 OcdRank 算法,支持数据增量更新,且时间复杂度低。

2.1.2　概率统计模型方法

由于基于排序的方法需要提前设置好社团的数目,存在不稳定性,为此,学者们提出使用概率统计模型进行社团划分。概率统计模型方法即指利用贝叶斯模型、先验概率、后验概率等方法计算节点属于社团的概率,从而达到划分目的。陈毅[33]提出多维度贝叶斯非参混合模型(MBNPM),对抓取到的每一维度的结构特征进行融合,利用聚类模型得到社团信息,该方法能够自动探索网络的社团数目并取得较优的社团划分效果。殷浩潇等基于混合网络的信息维统计量提出 DirCom 方法,对混合网络进行信息维上卷后,学习信息维的狄利克雷分布参数来表征某个社区,利用最大后验

chinaXiv:202204.00047v1

ChinaXiv合作期刊

概率实现社团划分[34]。S. Sengupta 和 Y. Chen 提出了针对随机块模型的混合网络谱聚类方法,应用了适用于大型网络的用于后验推理的变分 EM 算法,允许不同类型的节点拥有多个成员关系[35],但该算法未解决重叠社区问题。

这些算法的研究对象只限包含异构关系的混合网络,但实际网络关系较为复杂,不仅包含不同类型节点间的异构关系,也包含同类型节点间的同构关系。针对这种网络,童浩等基于 RankClus 算法将排名聚类方法与协同聚类方法相结合提出 RankCoClus 算法,选取论文、作者、术语、会议4种节点及会议-作者、作者-作者2种关系,实验证明其有效性[36]。R. Wang 等提出 ComClus 算法,该算法采用带自循环的星型模式来组织混合网络,并使用概率模型来表示对象的生成概率[37],实验表明,该方法的聚类效果更优。

从以上分析可见,基于排序的方法虽然时间复杂度较低,可以实现动态网络的社团划分,但其需要根据先验知识指定社团数目,当网络规模较大时,很难准确地进行预测,从而导致结果的不稳定性。概率统计模型方法不需要先验知识,稳定性较强,适用于大型网络的社团划分,但未解决重叠社区问题。

2.2 基于元路径的方法

多种类型的节点由多条链路连接而成,连接不同节点的链路都蕴含着不同的语义,这样的链路形成元路径[38]。元路径是一种有效的语义捕获工具,可以捕捉混合网络内丰富的语义信息[39-40],是混合网络的独特特征,也是一种特征提取方法[41]。因此,在多节点多关系混合网络中,基于元路径的社团划分方法相继涌现。

PathSim[40]是最早提出的基于元路径的算法,该算法针对同构网络提出,对于度量相同类型节点间的相似度表现较好。J. Li 等指出,大多数基于元路径的混合网络社团划分方法存在两个问题:①由元路径直接获得的相似度通常是一个偏差度量;②如何对不同元路径的相似性进行融合[42]。为此,他们基于 PathSim 的标准化来消除相似性偏差,设计了一种灵活的融合机制来动态优化结果,使社团划分结果更优。C. Shi 等[43]基于元路径提出一种可以度量相同或不同类型节点的相似性算法——HeteSim,该算法通过双向随机游走来计算相似性,在查询和聚类任务中表现优于传统算法,但是 HeteSim 只适用于单条元路径环境下,不能够捕获混合网络中的多种语义信息[44],且该算法复杂度高,不适合大规模网络。随后,X. F. Meng 等[45]提出了一种基于给定元路径和反向元路径的双随机游走过程来计算两个对象的相似性算法——AvgSim,其能够在大规模网络中应用,且聚类效果佳。

不同的元路径包含的信息不同,选择不同的元路径会导致不同的社团划分结果,如何在多条元路径中确定选取的元路径条数或者最优元路径是一难题。Y. Z. Sun 等[46]提出 PathSelClus 算法,它能够为混合网络中不同元路径分配不同的权重。吴瑶等提出一种多元图融合的混合网络嵌入方法,可以自动学习网络中的关键元路径[47]。C. Shi[41]等引入基于元路径的随机游走方法 HRank 来评估节点和元路径的重要性,实验结果显示了元路径的独特优势。

从以上分析可见,混合网络中基于元路径的社团划分方法大多是由同构网络 PathSim 方法改进得来。基于元路径的方法较为简便易懂,多条元路径能够捕获到混合网络中的丰富信息,但其算法复杂度较高,得到的相似度通常是一个偏差度量[42],对大规模网络的适用性较差。此外,不同的元路径包含的信息不同,如何准确地计算节点之间的相似度以展现出丰富的语义关联关系、如何在多条元路径中选择最优的元路径从而获得最优划分效果仍然是难题。

2.3 基于种子节点的方法

以种子为中心的方法成为了社团划分算法的一种新兴趋势[48],基于种子节点的方法的基本思想是识别网络中的某些特定节点,称为种子节点,再围绕这些节点构建社区[49-51]。

Z. Yakoubi 等首先提出种子节点驱动的社团划分算法 Licod,其基本思想是选择比大多数直接邻居具有更高中心性的节点作为种子节点,围绕这些节点进行本地社团计算,再从本地社团集合中进行划分[52],但该方法只适用于同构网络。M. Hmimida 等[48]将 Licod 算法扩展到混合网络中,称为 mux-Licod,该方法考虑了混合网络不同层节点之间的不同类型关系,实验结果表明该方法具有较好的实用性。薛维佳提出基于种子节点聚类的社团划分算法 NS-Clus,根据节点重要度以及二阶邻居选取种子节点,随后通过相似性度量对种子节点进行初始社团划分,并利用节点隶属于社团的概率将非种子节点加入到社团中,得到最终划分结果,在 DBLP 数据集以及 ACM 数据集上的测试结果表明了该算法的有效性[38]。

基于种子节点的方法是一种局部计算方法,该方法便于理解,适合处理大规模网络和动态网络[52]。但是如何高效地选择有效的种子节点仍未达成共识,并

chinaXiv:202304.00470v1

且在对非种子节点社团进行合并时,会出现大社区合并过度、小社区数量过多的问题。

2.4 扩展模块度算法

模块度最先是用于评价社团划分结果的指标,随着研究的深入,出现了基于模块度的社团划分算法[6,53-54],扩展模块度算法即是将适用于同构网络的模块度算法扩展到混合网络中。

M. E. J. Newman 等最先提出模块度优化算法 FN,该方法将每个节点看作一个社团,计算两两社团结合后的模块度值,采取模块度值增加最大或减少最小的社团结合方式,迭代直至模块度不再增加完成社团划分[6],但只适用于单节点类型的网络。R. Guimerà 等提出了适用于二分网络的扩展模块度算法,该算法能够独立地识别具有相似输出连接的节点和具有相似输入连接的节点[55],但其不具有普适性。T. Murata 等[56]提出了适用于 k 核网络的模块度算法,该法存在一般模块度算法都存在的分辨率限制问题,且不适用于一般形态的混合网络。X. Liu 等[57]提出复合模块度方法,其核心思想是将混合网络分解为多个子网络,对每个子网络中的模块度进行集成,基于 Louvain 算法优化复合模块度,实现社团划分,该方法不需先验知识,且适用于大规模网络与一般形态网络。

显而易见,扩展模块度算法由同构网络中的算法演变而来,稳定性较高,适用于大规模网络;但其网络适用性较差,仍避免不了模块度最大化的局限——分辨率限制,无法检测出大规模网络中的小社区[5,58-59]。

2.5 混合网络同构方法

由于同构网络的社团划分方法相对成熟,可以将混合网络降维成同构网络,再使用同构网络社团划分方法进行划分。降维方法主要有非负矩阵分解(NMF)[60]、主题模型[14]、主成分分析(PCA)[61]、线性判别分析(LDA)[62]等。

2.5.1 非负矩阵分解方法

非负矩阵分解方法对于任意给定的一个非负矩阵,都能分解为两个非负矩阵[63],分别为基矩阵和系数矩阵,利用系数矩阵来代替原矩阵实现降维。S. Tafavogh 提出一种基于矩阵分解和语义路径的混合网络社团划分方法[64],实验表明其有效性。X. C. Zhang 提出了一种非负矩阵三因子分解方法 HMFClus,利用相似性正则化将同类型对象之间的信息集成到 HMFClus 中,该方法可以同时对混合网络中所有类型的对象进行聚类[65]。黄瑞阳等利用多关系相似度矩阵融合动态混合网络中的信息,结合非负矩阵分解模型发现网络中的社团结构,该算法在社团划分上有效,但复杂度高[66]。J. Liu 等针对多层属性网络,从矩阵分解的角度提出了一种惩罚替代因子分解(PAF)算法来解决相应的优化问题,PAF 算法不仅社团划分效果好,且对网络形态的适用性强[67]。

2.5.2 主题模型方法

引入主题模型,可以挖掘出文本信息中隐藏的主题信息以提高社团划分的效果[68]。Q. Z. Mei 等充分利用统计主题模型和离散正则化的优点,通过正则化改进主题模型,实现社团划分[69]。王婷提出基于主题感知的 LDA-light 算法,将混合网络降维成同构网络或者二分网络,利用标签传播方法进行社团划分,该方法划分出来的社团带有语义信息,且普适性强,可以推广应用到实际场景中[70]。

2.5.3 主成分分析(PCA)与线性判别分析(LDA)法

这两种方法均属于线性降维方法,使用线性投影的方法将高维度数据映射到低维空间,其不同点在于,前者确保降维后的数据保留较多的原始信息,后者是使降维后的数据更易被区分[71]。现有研究只将这两种方法用于单节点类型的网络中[72-74]或二分网络中[75]。

混合网络同构方法虽然便于理解,但将混合网络降维成同构网络的过程复杂,易造成信息失真。非负矩阵分解方法的网络适用性强,但实现复杂度过高;主题模型方法利用语义信息进行社团划分,其结果更加可靠,且普适性较强;主成分分析与线性判别分析方法的网络适用性较差。

现如今越来越多的研究不局限于一种社团划分方法,多种方法的融合会使得社团划分效果更优。高苌婕等利用基于语义的元路径模型计算节点间的相似性,通过最小化目标函数值得到社团划分结果[77]。陈长赓提出了基于元路径计算相似性的标签传播算法(PathLPA)[76],并将其应用到 DBLP 混合网络中对作者节点进行社团划分,取得良好划分效果。张正林提出一种基于元路径抽取与种子社区的重叠社团划分算法 Hete_MESC,用户根据需求选取中心节点,从网络中抽取出关于中心节点的多路网络后对其进行社团划分,将划分结果作为种子社团,根据其他类型节点与种子社团之间的相似度最终实现所有节点的社团划分[26],该算法适用于任何形态的网络,且复杂度低。

综上所述,现有的针对多节点多关系混合网络的社团划分方法,大多是基于概率模型和元路径,基于种子节点、扩展模块度、混合网络同构方法仍处于探索阶

chinaXiv:202304.00470v1

ChinaXiv合作期刊

段。此外，各类方法仍存在不少问题亟待解决，可见对多节点多关系混合网络社团划分方法开展进一步研究还有很大空间。

3 社团划分效果常用的评估指标

社团划分效果的评估指标有很多种，对于不同的实验需求使用的评估指标也不一样。本文主要介绍在社团划分研究领域内使用最为广泛的 3 种指标：标准化互信息 NMI[76]、调整兰德指数 ARI[78] 和模块度 Q[6]，其中 NMI、ARI 是针对已知真实社团划分结果的评估指标，模块度 Q 是针对不知真实社团划分结果的评估指标，其对比如表 2 所示：

表 2 3 种社团划分评估指标对比

指标名称	是否已知真实社团划分结果	指标类型	取值范围	值与社团划分结果的关系
标准化互信息 NMI	是	衡量数据分布间的差异	[0,1]	正相关
调整兰德指数 ARI	是	衡量数据分布间的吻合程度	[-1,1]	正相关
模块度 Q	否	衡量社区强度	[0,1]	正相关

标准化互信息（NMI）是一种在信息论、概率论知识基础上产生的评估社区划分结果的相似性度量方法[76]，通常用于检测真实划分结果与实际划分结果之间的差异，可以直观地表现出社团划分结果的好坏。NMI 计算公式如下：

$$NMI=\frac{-2\sum_{i=1}^{C_A}\sum_{j=1}^{C_B}N_{ij}\times\log(\frac{N_{ij}\times N}{N_i\times N_j})}{\sum_{i=1}^{C_A}N_i\times\log(\frac{N_i}{N})+\sum_{j=1}^{C_B}N_j\times\log(\frac{N_j}{N})} \quad 式(1)$$

其中，A 和 B 为网络中划分出来的结果集，N 是所有节点的数量，C_A、C_B 分别代表 A、B 中社团的个数，N_{ij} 表示两个社团共有节点的个数，$N_i(N_j)$ 为 N 中第 i(j) 行元素之和。NMI 取值范围为[0,1]，值越大说明社团划分越准确。

ARI 从广义角度来讲，衡量的是两个数据分布的吻合程度，即每个点对在不同的社团划分下是否保持一致来比较社团划分结果与真实划分的相似性[79]，其定义如下：

$$ARI=a_{11}-\frac{\frac{(a_{11}+a_{01})(a_{11}+a_{10})}{a_{00}}}{\frac{(a_{11}+a_{01})+(a_{11}+a_{10})}{2}-(\frac{(a_{11}+a_{01})(a_{11}+a_{10})}{a_{00}}} \quad 式(2)$$

其中 a_{11} 表示在真实社团划分与实际社团划分中都属于同一社团的点对数，a_{00} 表示在真实社团划分与实际社团划分中都不属于同一社团的点对数，a_{10} 表示在真实社团中属于同一社团而在实际社团划分中不属于同一社团的点对数，a_{01} 表示在真实社团中不属于同一社团而在实际社团划分中属于同一社团的点对数[79]。其取值范围为[-1,1]，值越大说明实际划分结果与真实划分结果越吻合，与 NMI 相比，ARI 有更高的区分度。

模块度函数 Q 是由 M. E. J Newman 和 M. Girvan 提出，通过优化模块度 Q 可以获得更优的社团划分结果，模块度 Q 可以使社团内部节点的联系更紧密，因此它是一种衡量社区强度的指标[38]，其定义如下：

$$Q=\frac{1}{2m}\sum_{ij}\left[A_{ij}-\frac{k_ik_j}{2m}\right]\delta(C_i,C_j) \quad 式(3)$$

其中，i 和 j 是任意两个节点，k_i、k_j 分别为节点 i、j 的度，m 为网络中的总边数。当两个节点直接相连时 $A_{ij}=1$，否则为 0；C_i、C_j 分别为节点 i、j 属于的社团，若两节点属于同一个社团，则 $\delta=1$，否则为 0。其取值范围为[0,1]，Q 值越大说明划分的社区结构越稳定，效果也越好。

标准化互信息 NMI、调整兰德指数 ARI 和模块度 Q 是评估社团划分效果最常用的指标，但其对多节点多关系混合网络的评价有效性仍有待系统且深入的论证。在混合网络社团划分效果的评价中，除使用这 3 种评估指标外，部分学者使用自定义的指标，比如利用论文关键词相关性、论文主题相关性、作者相关性评价社团划分的效果[26]。可见，该领域仍缺乏一个标准统一的评估指标，该评估指标应该同时考虑同类型节点及不同类型节点间的连接强度，因此构建一个适用于该领域的社团划分效果评估指标是未来努力的方向。

4 多节点多关系混合网络社团划分的应用

多节点多关系混合网络的社团划分研究不仅具有理论意义，在实际应用中也存在着可行性和有效性。研究者们将社团划分方法应用到各个领域来发现社团结构以解决实际问题。本文选取社交媒体、学术网络、欺诈检测这 3 个常用领域，对研究者们在各领域中常用的社团划分方法进行阐述。

4.1 社交媒体

社交媒体网络的迅速发展使得其节点众多、关系

chinaXiv:202304.00470v1

错综复杂,对其进行社团划分在好友推荐、舆情监测等方面都具有现实意义,并且可以从网络层面了解各个社团并将它们与现实生活相关联[80]。社交网络中的一个关键任务就是推荐系统,而社团划分的任务就是对志同道合的人进行划分。概率生成模型方法是基于社团结构的推荐系统中最常用的方法,通过识别相似用户、根据用户的共同特征作精准推荐,该方法可以优化协同过滤方法存在的数据过载、推荐效率低等[81]问题。陈毅将贝叶斯非参混合模型 BNPM 方法应用到好友推荐中[33],与传统好友推荐算法相比取得较优的效果,同时提高了推荐效率。

4.2 学术网络

随着对科学结构研究的逐渐深入,构建关于作者、文献、关键词等节点和作者合作、文献引用、文献关键词隶属等关系的混合网络,并进行社团划分,可以了解更多的学者结构信息、展示不同角度的社团结构,为全面清晰地揭示科学共同体、科研结构、某一学科的发展脉络提供依据,这也成为学术网络研究领域的一个新视角。很多学者利用概率生成模型、元路径、种子节点方法在 DBLP 学术文献数据集上应用他们提出的多节点多关系混合网络社团划分算法[38,82-83],以验证算法的有效性。张正林构建了包含论文、作者、关键词、期刊 4 种节点以及论文引用、论文-作者著作、论文-关键词包含、论文-期刊发表 4 种关系的混合网络,基于元路径抽取和种子节点的方法进行社团划分后,对比作者社团和论文社团,发现“论文社团规模较小,研究领域单一;作者社团规模较大,研究领域分散”的特点[26]。

ChinaXiv:202304.00407v1

4.3 欺诈检测

欺诈检测在电信网络、医疗保健等现实生活中有着巨大的应用。在各类欺诈检测中,均涉及节点众多、数据量大且分布不均的问题,传统的异常检测方法很难检测出异常,而多节点多关系混合网络的社团划分方法在有效简化问题的同时能够更多地关注节点间的关系,为欺诈检测提供了新的方向。扩展模块度算法是该场景下最常使用的方法[84-85],栾婷婷将普通住院数据中的医生和药品建模为混合加权网络,提出模块度优化算法 FNO 将医生和药品划分到相应的社区,最后再通过医生和药品社团的对比,发现异常医生,实现医疗保险领域的欺诈问题检测[84]。

5 讨论与展望

本文围绕多节点多关系混合网络的相关方法研究,梳理了其社团划分方法,介绍了常用的社团划分评估指标及应用场景。目前对混合网络社团划分方法的研究还处于起步阶段,经过梳理发现多节点多关系混合网络的社团划分方法主要有基于概率生成模型、基于元路径、基于种子节点、扩展模块度、混合网络同构等,本文对比了各类方法的适用性和特点,指出构建一个网络适用性强、复杂度低、同时结合网络拓扑结构与文本信息的社团划分算法很有必要。现如今常用的标准化互信息 NMI、调整兰德指数 ARI 和模块度 Q 三种社团划分效果评估指标,均有各自特点及适用范围,但其对多节点多关系混合网络的社团划分效果的评价有效性仍有待系统且深入的论证,提出适用于该领域的统一评估指标也是今后的研究方向之一。

多节点多关系混合网络打破了传统同构网络的单一局限性,对其进行分析可以挖掘出隐藏的丰富信息,但其特性使得社团划分算法面临了不少挑战:①网络具有多种类型的节点与关系,如何融合多层网络、合理有效地利用混合网络中的拓扑结构信息和节点属性信息是面临的首要问题[28];②网络规模大,现实网络节点数量众多且其之间关系稀疏,设计出一个适用于大规模网络且划分效果好的算法面临更大的困难;③存在一定量的无连接的同类型节点或关系,不利于相似度度量的计算;④目前对重叠社区进行识别的算法并不多,然而在实际网络中,一个节点很有可能同时属于多个社区,需要利用有效的算法对其进行区分。这些都致使研究异构网络的社团划分算法十分具有挑战性[26],也是今后研究需要解决的难题。

此外,目前多节点多关系混合网络的社团划分方法在社交媒体、学术网络、欺诈检测场景下的应用大多是针对静态网络的,未考虑到数据集的变化带来的影响,在未来研究中,如何构建一个适用于动态网络的混合网络社团划分算法,以深度揭示科学结构的动态变化,值得进一步研究。对于多节点多关系混合网络的前沿研究不仅局限于社团划分上,还有链接预测、节点分类、语义搜索等任务,这对于混合网络的研究很有现实意义,也是今后混合网络研究的方向之一。

参考文献:

[1] NEWMAN M E J. Networks[M]. Oxford: Oxford University Press, 2018.
[2] WASSERMAN S, GALASKIEWICZ J. Advances in social network analysis: research in the social and behavioral sciences[M]. Los Angeles: Sage, 1994.
[3] BADER G D, HOGUE C W. An automated method for finding molecular complexes in large protein interaction networks[J]. BMC

ChinaXiv合作期刊

bioinform, 2003, 4(2): 1 – 27.
[4] SPORNS O. Networks of the brain[M]. Cambridge: MIT Press, 2010.
[5] FORTUNATO S. Community detection in graphs[J]. Physics reports, 2009, 486(3): 75 – 174.
[6] NEWMAN M E J, GIRVAN M. Finding and evaluating community structure in networks[J]. Physical review e, 2004, 69(2): 026113.
[7] DUCH J, ARENAS A. Community detection in complex networks using extremal optimization[J]. Physical review e statistical nonlinear & soft matter physics, 2005, 72(2): 027104.
[8] SCHLITT T, BRAZMA A. Current approaches to gene regulatory network modelling[J]. BMC bioinform, 2007, 8(6):1 – 22.
[9] 郑玉艳, 王明省, 石川, 等. 异质信息网络中基于元路径的社团发现算法研究[J]. 中文信息学报, 2018, 32(9): 132 – 142.
[10] PORTER M A, ONNELA J P, MUCHA P J. Communities in networks[J]. Notices of the American Mathematical Society, 2009, 56(9): 4294 – 4303.
[11] JAVED M A, YOUNIS M S, LATIF S, et al. Community detection in networks: a multidisciplinary review[J]. Journal of network and computer applications, 2018, 108(4): 87 – 111.
[12] KARATAS A, SAHIN S. Application areas of community detection: a review[C]// 2018 International congress on big data, deep learning and fighting cyber terrorism. Ankara: IEEE. 2018, 65 – 70.
[13] 李辉, 陈福才, 张建朋, 等. 复杂网络中的社团发现算法综述[J]. 计算机应用研究, 2021, 38(6): 1611 – 1618.
[14] 张瑞红, 陈云伟, 邓勇. 用于科学结构分析的混合网络社团划分方法述评[J]. 图书情报工作, 2019, 63(4): 135 – 141.
[15] 钱学森, 于景元, 戴汝为. 一个科学新领域——开放的复杂巨系统及其方法论[J]. 自然杂志, 1990, 13(1): 3 – 10.
[16] MCCORD M R. Urban transportation networks: equilibrium analysis with mathematical programming methods[J]. Transportation research part a general, 1987, 21(6): 481 – 484.
[17] 孙艺洲, 韩家炜. 异构信息网络挖掘:原理与方法[M]. 北京: 机械工业出版社, 2016.
[18] 霍朝光, 张斌, 董克. 复杂网络视域下的学术行为预测研究述评: 选题、合作与引用[J]. 情报理论与实践, 2021, 44(6): 180 – 188,27.
[19] RAN W, SHI C, YU P S, et al. Integrating clustering and ranking on hybrid heterogeneous information network[C]// Pacific-Asia conference on knowledge discovery and data mining. Berlin: Springer, 2013: 583 – 594.
[20] VAN DEN BESSELAAR P, HEIMERIKS G. Mapping research topics using word-reference co-occurrences: a method and an exploratory case study[J]. Scientometrics, 2006, 68(3): 377 – 393.
[21] HAN J W. Mining heterogeneous information networks: principles and methodologies[J]. ACM sigkdd explorations newsletter. 2012, 3(2): 1 – 159.
[22] 康宇航. 基于“耦合 – 共引”混合网络的技术机会分析[J]. 情报学报, 2017(2): 170 – 179.
[23] BERLINGERIO M, COSCIA M, GIANNOTTI F. Finding and characterizing communities in multidimensional networks[J]. 2011 international conference on advance in social networks analysis and mining, 2011(8): 490 – 494.
[24] SUTHERS D, FUSCO J, SCHANK P, et al. Discovery of community structures in a heterogeneous professional online network[J]. 2013 46th Hawaii international conference on system sciences, 2013(3): 3262 – 3271.
[25] TANG L, LIU H. Community detection and mining in social media [J]. Community detection and mining in social media, 2010, 2 (1): 1 – 137.
[26] 张正林. 大规模异构信息网络社区发现算法与社区特征研究[D]. 北京: 北京邮电大学, 2017.
[27] SUN Y Z, HAN J W, ZHAO P X, et al. RankClus: integrating clustering with ranking for heterogeneous information network analysis[C]//ACM sigkdd international conference on knowledge discovery & data mining. New York: ACM Press, 2009: 565 – 576.
[28] SUN Y Z, HAN J W. Ranking-based clustering of heterogeneous information networks with star network schema[C]// ACM sigkdd international conference on knowledge discovery & data mining. New York: ACM Press, 2009: 797 – 806.
[29] JI M, HAN J W, DANILEVSKY M. Ranking-based classification of heterogeneous information networks[C]// ACM sigkdd international conference on knowledge discovery & data mining. New York: ACM Press, 2011: 1298 – 1306.
[30] 赵焕. 基于异构网络聚类的 Web 服务推荐系统研究[D]. 重庆: 重庆大学,2015.
[31] GUPTA M, AGGARWAL C C, HAN J, et al. Evolutionary clustering and analysis of bibliographic networks[C]// International conference on advances in social networks analysis and mining. Taiwan: IEEE, 2011: 63 – 70.
[32] QIU C H, CHEN W, WANG T J, et al. Overlapping community detection in directed heterogeneous social network[J]. Web-age information management, 2015(6): 490 – 493.
[33] 陈毅. 基于统计推理的复杂网络社区结构分析[D]. 哈尔滨: 哈尔滨工业大学, 2016.
[34] 殷浩潇, 李川. 异构信息网络概率模型研究及社区发现算法[J]. 现代计算机(专业版), 2016(3): 3 – 6.
[35] SENGUPTA S, CHEN Y. Spectral clustering in heterogeneous networks[J]. Statistica sinica, 2015, 25(3): 1081 – 1106.
[36] 童浩,余春艳. 基于排名分布的异构信息网络协同聚类算法[J]. 小型微型计算机系统, 2014, 35(11): 2445 – 2449.
[37] WANG R, SHI C, YU P S, et al. Integrating clustering and ranking on hybrid heterogeneous information network[C]// Pacific-Asia conference on knowledge discovery and data mining. Berlin: Springer, 2013: 583 – 594.
[38] 薛维佳. 异构信息网络中基于聚类的社区发现方法研究[D]. 包头:内蒙古科技大学, 2020.
[39] SUN Y Z, HAN J W, YAN X F, et al. Pathsim: meta path-based

chinaXiv:202304.00470v1

top-k similarity search in heterogeneous information networks[J]. Proceedings of the VLDB endowment, 2011, 4(11): 992－1003.

[40] SHI C, LI Y T, ZHANG J W, et al. A survey of heterogeneous information network analysis[J]. IEEE transactions on knowledge and data engineering, 2017, 29(1): 17－37.

[41] SHI C, YU P S. Heterogeneous information network analysis and applications[M]. Switzerland: Springer International Publishing, 2017:.

[42] LI J, SUN P Y, MAO Q R, et al. Path-Graph fusion based community detection over heterogeneous information network[C]// 2018 IEEE 20th international conference on high performance computing and communications. New Jersey: IEEE, 2018: 274－281.

[43] SHI C, KONG X N, YU P S, et al. Relevance search in heterogeneous networks[C]// Proceedings of the 15th international conference on extending database technology. New York: ACM Press, 2012: 180－191.

[44] 丁平尖. 基于元路径的异构信息网络挖掘方法研究[D]. 长沙:湖南大学, 2015.

[45] MENG X F, SHI C, LI Y T, et al. Relevance measure in large-scale heterogeneous networks[C]// Asia-Pacific Web conference. Berlin: Springer, 2014: 636－643.

[46] SUN Y Z, NORICK B, HAN J W, et al. Pathselclus: integrating meta-path selection with user-guided object clustering in heterogeneous information networks[J]. ACM transactions on knowledge discovery from data, 2013, 7(3): 1－23.

[47] 吴瑶, 申德荣, 寇月, 等. 多元图融合的异构信息网嵌入[J]. 计算机研究与发展, 2020, 57(9): 1928－1938.

[48] HMIMIDA M, KANAWATI R. Community detection in multiplex networks: a seed-centric approach[J]. Networks & heterogeneous media, 2015, 10(1): 71－85.

[49] KANAWATI R. LICOD: leaders identification for community detection in complex networks[C]//2011 IEEE third international conference on privacy, security, risk and trust and 2011 IEEE third international conference on social computing. New Jersey: IEEE, 2011: 577－582.

[50] PAPADOPOULOS S, KOMPATSIARIS Y, VAKALI A. A graph-based clustering scheme for identifying related tags in folksonomies[C]// Proceedings of the 12th international conference on data warehousing and knowledge discovery. Berlin: Springer, 2010: 65－76.

[51] SHAH D, ZAMAN T. Community detection in networks: the leader-follower algorithm[J]. Workshop on networks across disciplines in theory and applications, 2010, 1050(2):1－8.

[52] YAKOUBI Z, KANAWATI R. LICOD: a Leader-driven algorithm for community detection in complex networks[J]. Vietnam journal of computer science, 2014, 1(4): 241－256.

[53] TANG L, WANG X F, LIU H. Uncoverning groups via heterogeneous interaction analysis[C]// Ninth IEEE international conference on data mining. New Jersey: IEEE Computer Society, 2009: 503－512.

[54] NICOSIA V, MANGIONI G, CARCHIOLO V, et al. Extending the definition of modularity to directed graphs with overlapping communities[J]. Journal of statistical mechanics: theory and experiment, 2009(3): 3166－3168.

[55] GUIMERÀ R, MARTA S P, AMARAL L A. Module identification in bipartite and directed networks[J]. Physical review e statistical nonlinear & soft matter physics, 2007, 76(2): 036102.

[56] MURATA T, IKEYA T. A new modularity for defecting one-to-many correspondence of communities in bipartite networks[J]. Advances in complex systems, 2010, 13(1): 19－31.

[57] LIU X, LIU W C, MURATA T, et al. A framework for community detection in heterogeneous multi-relational networks[J]. Advances in complex systems, 2014, 17(6): 1450018.

[58] LANCICHINETTI A, FORTUNATO S. Limits of modularity maximization in community detection[J]. Physical review e statistical nonlinear & soft matter physics, 2011, 84(6): 066122.

[59] FORTUNATO S, BARTHÉLEMY M. Resolution limit in community detection[J]. Proceedings of the National Academy of Sciences of the United States of America, 2007, 104(1): 36－41.

[60] LEE D D, SEUNG H S. Learning the parts of objects by non-negative matrix factorization[J]. Nature, 1999, 401(6755): 788－791.

[61] JOLLIFFE I T. Principal component analysis[M]. Berlin: Springer-Verlag, 2002.

[62] SCHOLKOPFT B, MULLERT K-R. Fisher discriminant analysis with kernels[J]. Neural neworks for signal processing ix, 1999, 1(1):41－48.

[63] LIU T L, GONG M M, TAO D C. Large-cone nonnegative matrix factorization[J]. IEEE transactions on neural networks and learning systems, 2016, 28(9): 2129－2142.

[64] TAFAVOGH S. Community detection on heterogeneous networks by multiple semantic-path clustering[C]// IEEE international conference on computational aspects of social networks. New Jersey: IEEE, 2014: 7－12.

[65] ZHANG X C, LI H X, LIANG W X, et al. Multi-type co-clustering of general heterogeneous information networks via nonnegative matrix tri-factorization[C]// 2016 IEEE 16th international conference on data mining. New Jersey: IEEE, 2016: 1353－1358.

[66] 黄瑞阳, 吴奇, 朱宇航. 基于联合矩阵分解的动态异质网络社区发现方法[J]. 计算机应用研究, 2017, 34(10): 2989－2992.

[67] LIU J, WANG J Z, LIU B H. Community detection of multi-layer attributed networks via penalized alternating factorization[J]. Mathematics, 2020, 8(2): 239－258.

[68] 刘培奇, 孙捷焓. 基于 LDA 主题模型的标签传递算法[J]. 计算机应用, 2012, 2(2): 403－406.

[69] MEI Q Z, CAI D, ZHANG D, et al. Topic modeling with network regularization[C]//Proceeding of the 17th international conference on World Wide Web. New York: ACM Press, 2008: 101－110.

chinaXiv:202304.00470v1

[70] 王婷. 异构社交网络中社区发现算法研究[D]. 北京: 中国矿业大学, 2016.
[71] 保丽红. 主成分分析与线性判别分析降维比较[J]. 统计学与应用, 2020, 9(1): 47－52.
[72] LIN L, XIA Z M, LI S H, et al. Detecting overlapping community structure via an improved spread algorithm based on pca[C]// International conference on computer science and software engineering. Lancaster: DEStec, 2014: 115－121.
[73] LI L, FAN K F, ZHANG Z Y, et al. Community detection algorithm based on local expansion K-means[J]. Neural network world, 2016, 26(6): 589－605.
[74] YUAN P Y, WANG W, SONG M Y. Detecting overlapping community structures with pca technology and member index[C]// Proceedings of the 9th EAI international conference on mobile multimedia communications. New York: ACM Press, 2016: 121－125.
[75] LIU W, CHEN L. Community detection in disease-gene network based on principal component analysis[J]. Tsinghua science and technology, 2013, 18(5): 454－461.
[76] 陈长赓. 异构信息网络下基于元路径的节点重要性度量和社区发现[D]. 昆明: 云南大学, 2019.
[77] 高苌婕, 彭敦陆. 面向DBWorld数据挖掘的学术社区发现算法[J]. 计算机应用研究, 2017(7): 2059－2062.
[78] SANTOS J M, EMBRECHTS M. On the use of the adjusted rand index as a metric for evaluating supervised classification[M]. Berlin: Springer, 2009.
[79] 王益文. 复杂网络节点影响力模型及其应用[D]. 杭州: 浙江大学, 2015.
[80] KARATAS A, SAHIN S. Application areas of community detection: a review[J]. 2018 international congress on big data, deep learning and fighting cyber terrorism (ibigdelft), 2019(1): 65－70.
[81] 张海涛, 周红磊, 张鑫蕊, 等. 在线社交网络的社区发现研究进展[J]. 图书情报工作, 2020, 64(9): 142－152.
[82] QIAO Y Q, NIU K, DU S, et al. Community detection analysis of heterogeneous network[J]. 2015 international conference on cyber-enabled distributed computing and knowledge discovery, 2015(10): 509－512.
[83] HUANG W H, LIU Y, CHEN Y G. Mixed membership stochastic blockmodels for heterogeneous networks[J]. Bayesian analysis, 2019, 15(3): 711－736.
[84] 栾婷婷. 基于异构网络社区划分的医疗滥用检测研究[D]. 济南: 山东大学, 2019.
[85] 刘殿中. 动态金融复杂数据的欺诈检测[D]. 青岛: 青岛大学, 2020.

作者贡献说明:
蒋璐:收集整理资料,撰写并修改论文;
陈云伟:提出综述撰写思路及修改意见。

A Review of Community Detection in Hybrid Networks with Multiple Nodes and Multiple Relationships

Jiang Lu[1,2]　Chen Yunwei[1,2]
[1] Scientometrics & Evaluation Research Center (SERC), Chengdu Library and Information Center of Chinese Academy of Sciences, Chengdu 610041
[2] Department of Library Information and Archives Management, School of Economics and Management, University of Chinese Academy of Sciences, Beijing 100190

Abstract: [**Purpose/significance**] By sorting out the community detection methods of multi-node and multi-relationship hybrid network, we can analyze the problems and difficulties existing in the community detection methods, and predict the development trend in the future. [**Method/process**] In this paper, the methods of multi-node type and multi-relation type hybrid network community detection were systematically reviewed, and were described from five aspects: based on probabilistic generation model, meta-path, seed node, expansion modularity and isomorphism of hybrid networks. This paper summarized the commonly used evaluation indicators for community detection in hybrid networks: Standardized Mutual Information(NMI), Adjusted Rand Index(ARI) and Modularity Q, and pointed out three application scenarios of social media, academic network and fraud detection. [**Result/conclusion**] This paper summarizes the applicability, advantages and disadvantages of the community detection methods of multi-node and multi-relationship hybrid network, reveals the challenges faced by the current development, provides a new perspective for the subsequent hybrid network analysis and research, and looks forward to the related research directions that may be further expanded in the future.

Keywords: multi-node and multi-relation　hybrid network　community　evaluation

chinaXiv:202304.00470v1